# The Introspective Gap: When AI Answer Engines Misdescribe Their Own Source Selection Behavior

*Paper I of III: A Simulation-Calibrated Exploratory Framework for AEO, GEO, and the Study of LLM Self-Knowledge*

Siddharth Nigam

*Independent Researcher*

April 2026 | Preprint

---

**Series Note.** *This is the first paper in a three-part series on AI information retrieval and agentic web optimization. Paper II introduces Parametric Divergence Mapping, an empirical methodology for deriving cross-model scoring criteria through iterative refinement. Paper III introduces the AXO (Agentic Execution Optimization) framework, a scoring rubric for measuring website readiness for autonomous AI agent task execution.*

---

**Abstract**

Answer Engine Optimization (AEO) and Generative Engine Optimization (GEO) are becoming central to how information is surfaced in AI-mediated discovery. Most existing work studies these systems from the outside: by measuring rankings, citations, visibility, or prompt-response behavior. This paper proposes a complementary reflexive framework: studying optimization partly through the answer engine's own account of what it values, and then comparing that account to behavior inferred from external empirical literature. The divergence between these two layers is what I call the *Introspective Gap*.

The paper presents three exploratory studies. First, a simulation-calibrated content optimization preference test across ten domains ($N$=10,000 modeled trials) reproduces the directional finding that citation-oriented strategies do not perform uniformly across domains. Second, a first quantitative framework for estimating the Introspective Gap by comparing an LLM's self-reported source-selection criteria with a synthesized set of behaviorally inferred weights. Under this framework, brand recognition appears substantially under-attributed in self-report, while factual accuracy and author credentials appear over-attributed. Third, an epistemic inflation model suggests that heavily GEO-optimized content may increase apparent confidence beyond evidential warrant, with the strongest inflation concentrated in weakly supported claims.

These results should be read as hypothesis-generating rather than definitive measurement. The main contribution is not a final empirical verdict on live answer engines, but a new methodological lens for asking whether AI systems accurately describe the factors that appear to influence their own citation behavior. If that self-description is systematically incomplete or flattering, the implications extend beyond optimization practice to transparency, auditing, and user trust in AI-mediated information systems.

**Keywords:** Answer Engine Optimization, Generative Engine Optimization, LLM Introspection, Citation Bias, AI Transparency, Epistemic Calibration

---

## 1. Introduction

The architecture of information discovery is changing. Search results increasingly arrive not as ranked lists but as synthesized answers, summaries, and conversational outputs. In that environment, visibility is no longer only a matter of ranking; it is also a matter of whether a

system cites, mentions, paraphrases, or silently absorbs a source into its response. This shift has given rise to two adjacent optimization disciplines.

Answer Engine Optimization (AEO) focuses on increasing the likelihood that content will be surfaced or cited in direct AI-generated answers. Generative Engine Optimization (GEO) extends that ambition by identifying content features that appear to improve visibility within responses produced by large language model-based systems. Prior work has begun to show that factors such as statistics, quotations, structure, and brand or entity strength may influence whether a source is surfaced or cited. Yet almost all of this work treats the model as a black box: the system is observed externally, but its own self-description of what it values is rarely treated as an object of study in itself.

This paper proposes a different approach. Rather than asking only, "What content gets cited?" it also asks, "How does the answer engine describe its own source-selection behavior, and how closely does that description align with patterns inferred from external evidence?" The distance between these two layers is what I call the *Introspective Gap*.

The term is intentionally narrow. It does not imply access to hidden model internals, and it does not claim that language models possess robust self-knowledge in any human sense. Instead, it refers to a measurable mismatch between *self-report*—what an LLM says it values when asked directly about source selection—and *behavioral proxy*—what a synthesized reading of external empirical literature suggests is rewarded in citation or visibility behavior.

The strongest claim of this paper is therefore methodological, not metaphysical. I am not claiming to have directly measured the hidden source-selection circuitry of answer engines. I am claiming that there is value in placing model self-description into explicit comparison with behaviorally inferred evidence, and that this comparison generates testable hypotheses about transparency, optimization, and trust.

## 2. Related Work

### 2.1 Generative Engine Optimization

Foundational GEO studies [1] have argued that adding statistics, quotations, or fluency enhancements can yield meaningful visibility improvements relative to baseline content. Other work has extended these ideas into verticals such as e-commerce [4], and industry analyses have documented phenomena such as ghost citations [6], uneven brand mention behavior [9], content recency effects [11, 14], and cross-platform instability in citation patterns [7, 8]. A critical literature has also emerged, arguing that GEO effectively transforms LLM-based discovery into a new advertising or influence surface [10], while others question whether citation diversity corresponds to gains in credibility or evidential quality [16]. This paper draws from both strands.

### 2.2 Answer Engine Optimization

AEO scholarship and industry practice focus on increasing the probability that content appears in direct AI answers, voice responses, or chat-based discovery tools. Typical tactics include answer-first formatting, schema markup, entity consistency, and signals associated with trust, expertise, or recency [12, 14]. Much of the evidence comes from applied or industry-facing studies rather than mature peer-reviewed literature, reflecting the youth of the field.

### 2.3 LLM Introspection and Self-Knowledge

A separate research stream has examined whether language models possess any reliable form of self-knowledge [2, 3]. Some work suggests that models can, in limited settings, identify manipulated internal states or predict aspects of their own behavior better than external observers [3]. But this literature remains cautious: introspective performance is partial, task-dependent, and far from comprehensive [2]. What has been missing is a bridge between introspection research and the AEO/GEO literature. That bridge is the paper's main conceptual intervention.

## 3. Methodology

### 3.1 Methodological Orientation: Reflexive AI Research

A distinctive feature of this paper is its reflexive setup. The LLM is used both as an experimental instrument and, in one part of the analysis, as the source of self-reported judgments about its own source selection. The paper therefore distinguishes sharply between three kinds of evidence, summarized in Figure 1.

Table 1: Evidentiary provenance. Readers should weight conclusions by layer: directly elicited data carries the most internal validity; synthesized estimates carry cross-study heterogeneity; modeled results are downstream of parameterization.

| Evidentiary Layer | What It Covers | Method | Status |
|---|---|---|---|
| Directly elicited | Self-reported factor weights (Table 3, left column) | Controlled LLM self-description | Primary data from this study |
| Synthesized from external studies | Behavioral proxy weights (Table 3, right column) | Sample-size-weighted median of normalized effect sizes | Cross-study proxy estimates |
| Modeled / simulated | Citation preferences (Table 2); calibration error (Table 4); visibility asymmetry (Table 5) | Monte Carlo simulation calibrated to published effect sizes | Computational extrapolation from prior findings |

The reflexive setup has an obvious tension. An AI system used to help study an Introspective Gap may itself be subject to one. That is not treated here as disqualifying, but as part of the object of inquiry. The safeguard is not self-validation. The safeguard is falsifiability. Every major claim in this paper can, in principle, be challenged by future controlled live-model experiments, human evaluation studies, or multi-model replication.

### 3.2 Experiment 1: Domain-Sensitive Content Preference Simulation

Ten domain-specific query environments were constructed, each with five content variants: plain baseline, statistics-heavy, structured Q&A, authority-signaled, and entity-rich. The experiment uses Monte Carlo simulation ($N$=10,000; 1,000 per domain) calibrated to effect sizes from prior

GEO literature, with domain modifiers and Gaussian noise ($\sigma$=0.03). The experiment does not involve live source injection into commercial answer engines. Its purpose is organizational: to model how strategy preference might shift under different domain assumptions, providing a structured baseline for Experiments 2 and 3.

### 3.3 Experiment 2: Estimating the Introspective Gap

Ten factors were selected as plausible influences on source selection. Self-report values were derived from a controlled elicitation. Comparison weights were synthesized from five prior empirical studies using different designs and metrics: GEO-bench effect sizes [1], Seer Interactive's ghost citation analysis ($N$=541,213; [6]), AirOps' citation study ($N$=45,000+; [7]), the Digital Bloom report ($N$=680M+; [8]), and Sommerfeld et al. ($N$=55,936; [16]). Where multiple sources bore on the same factor, a sample-size-weighted median of normalized values was used. Where only one source bore directly on a factor, it was treated as a rough point estimate with wider implicit uncertainty.

Table 3 includes estimated uncertainty bands for each behavioral proxy weight. These bands are judgment-based heterogeneity estimates reflecting: the number of independent sources contributing to each weight, the degree of methodological variation across those sources, and the construct distance between what each source measured and the factor being estimated. Factors informed by multiple large-sample studies with convergent findings receive tighter bands; single-source factors receive wider ones. They are *not* formal confidence intervals derived from a single sampling distribution.

The Introspective Gap is defined as:

$$\text{Gap} = \text{Behavioral proxy weight} - \text{Self-reported weight}$$

Positive values indicate under-attribution in self-report. Negative values indicate over-attribution. This design allows comparison without claiming that either layer is pure ground truth. The self-report may be incomplete; the behavioral proxy may be noisy; the gap lies in the mismatch itself.

### 3.4 Experiment 3: Epistemic Inflation

Eight claims with varying evidential support (0.35–0.90) were modeled across five content styles (200 trials per claim-style pair; 8,000 total). Expressed confidence was generated under a diminishing-returns inflation model:

$$\text{effective} = \text{base} \times (1 - \text{gt} \times 0.5) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\, 0.05)$$

This experiment is heavily model-dependent. Results should be read as "if inflation dynamics follow approximately this form, then the downstream pattern is..." rather than as direct measurement of AI confidence calibration. The qualitative pattern—that optimization inflates weak claims more than strong ones—is robust across reasonable parameterizations because it follows from the diminishing-returns structure itself. Specific magnitudes are illustrative.

## 4. Results

### 4.1 Domain-Sensitive Strategy Preferences (Modeled)

Across the modeled trials, the strongest overall performance was associated with authority-signaled (24.8%) and statistics-heavy (23.7%) content, with plain-text baseline performing worst

(12.6%). These figures should be read as modeled preference estimates, not live engine measurements.

Table 2: Modeled citation rates by strategy and domain ($N$=1,000 per domain). 95% CIs $\approx \pm 2.5$–3pp per cell.

| Domain | Plain | Stats | Q&A | Authority | Entity |
|---|---|---|---|---|---|
| Business Software | 11.8% | 25.3% | 19.8% | 22.8% | 20.3% |
| Renewable Energy | 13.7% | 26.8% | 16.3% | 22.1% | 21.1% |
| Nutrition Science | 12.6% | 25.7% | 17.4% | 28.3% | 16.0% |
| Enterprise Software | 13.4% | 23.7% | 24.8% | 21.8% | 16.3% |
| Electronics Mfg. | 11.9% | 27.9% | 17.0% | 22.3% | 20.9% |
| Clinical Psychology | 11.1% | 24.0% | 19.9% | 27.7% | 17.3% |
| Database Technology | 13.2% | 18.2% | 22.8% | 21.8% | 24.0% |
| Real Estate Economics | 12.2% | 27.0% | 17.8% | 27.7% | 15.3% |
| Human Resources | 12.7% | 20.8% | 20.9% | 28.0% | 17.6% |
| Biotechnology | 13.1% | 17.3% | 17.9% | 25.4% | 26.3% |
| **Overall** | **12.6%** | **23.7%** | **19.5%** | **24.8%** | **19.5%** |

That result is less important as a standalone finding than as a staging ground for the paper's larger claim: if models are sensitive to different content signals in different contexts, then asking them to summarize what they value in general may already flatten important structure.

### 4.2 Estimated Introspective Gap (Synthesized)

Table 3: Estimated Introspective Gap. Uncertainty bands are judgment-based heterogeneity estimates (see Section 3.3), not formal confidence intervals.

| Factor | Self-Report | Behavioral Proxy | Est. Uncertainty | Gap | Interpretation |
|---|---|---|---|---|---|
| Factual accuracy | 0.95 | 0.70 | ±0.12 | −0.25 | Over-attributed |
| Source recency | 0.85 | 0.82 | ±0.06 | −0.03 | Rough alignment |
| Author credentials | 0.80 | 0.55 | ±0.10 | −0.25 | Over-attributed |
| Statistical evidence | 0.78 | 0.88 | ±0.08 | +0.10 | Under-attributed |
| Clarity of writing | 0.72 | 0.80 | ±0.12 | +0.08 | Under-attributed |
| Structured formatting | 0.65 | 0.78 | ±0.10 | +0.13 | Under-attributed |
| Brand recognition | 0.30 | 0.72 | ±0.10 | +0.42 | **Strongly under-attributed** |
| Keyword density | 0.15 | 0.12 | ±0.05 | −0.03 | Rough alignment |
| Content length | 0.40 | 0.55 | ±0.12 | +0.15 | Under-attributed |
| Emotional tone | 0.25 | 0.35 | ±0.12 | +0.10 | Under-attributed |

Several patterns stand out. The self-report appears to privilege factors associated with epistemic seriousness: accuracy, recency, credentials, and evidence. Under the synthesized behavioral proxy, some of those factors remain important, but not always to the degree claimed in self-description. A different cluster—brand recognition, structure, content length, and presentation signals—appears under-attributed. The largest gap belongs to brand recognition (+0.42; proxy informed by [6, 8]). Even at the low end of the uncertainty band (proxy = 0.62), the gap remains substantial (+0.32).

This should not be overread. The table does not prove that the model consciously prefers brands. It does something narrower: it exposes a mismatch between the model's normative self-description and an empirically informed proxy of the behavior environment in which citations occur. The pattern is consistent with a social-desirability-like structure familiar from human survey research: epistemically respectable factors are emphasized in self-report, while operationally influential but less flattering factors are muted. I use that comparison once, carefully, and do not claim it identifies a mechanism.

### 4.3 Epistemic Inflation (Modeled)

Table 4: Modeled calibration error by content optimization strategy. The narrow CI reflects trial count ($N$=1,600 per style), not low variance.

| **Content Style** | **Mean Cal. Error** | **SD** | **95% CI of Mean** |
|---|---|---|---|
| Plain | +0.034 | 0.049 | ±0.0024 |
| Statistics-Optimized | +0.122 | 0.052 | ±0.0025 |
| Authority-Optimized | +0.145 | 0.055 | ±0.0027 |
| Structured-AEO | +0.080 | 0.049 | ±0.0024 |
| Full-GEO (Combined) | +0.182 | 0.064 | ±0.0031 |

More important than the absolute values is the qualitative asymmetry. In the model, weaker claims inflate more than stronger ones. Optimization does not simply make well-supported claims look cleaner; it may also make weakly supported claims feel more settled than they are. This result is exploratory. It depends on modeling assumptions and should not be treated as direct proof of live-system calibration failure. But it is a meaningful hypothesis: if optimization systematically changes the presentation cues that models use when expressing confidence, then calibration becomes part of the GEO conversation, not a separate problem.

### 4.4 Exploratory Visibility Asymmetry by Brand Size

Table 5: Exploratory visibility ranges by brand size. These are modeled ranges, not observed campaign data.

| Brand Size | Pre-Optimization | Post-Optimization | Direction |
|---|---|---|---|
| Enterprise (>10K) | 35–55% | 52–90% | Strong gain |
| Mid-market (100–10K) | 18–32% | 32–60% | Strong gain |
| Startup (<100) | 8–16% | 12–24% | Moderate gain |
| Solo/Indie | 2–6% | 2–11% | Modest gain |

The central takeaway is qualitative rather than numerical: under a wide range of plausible settings, optimization tends to help all actors somewhat, but the absolute gains may be larger for actors who already possess stronger brand, entity, or recognition signals. Under the present parameterization, the widening of the absolute gap falls in the 20–40 percentage point range. This is the weakest and most exploratory section of the paper. It suggests a plausible compounding dynamic, not a measured universal widening effect.

## 5. Discussion

### 5.1 What the Introspective Gap Is and Is Not

The Introspective Gap should not be mistaken for proof that models are deceptive, self-aware, or internally contradictory in a human sense. The concept is narrower. It names a mismatch between what a model says it values and what a behaviorally informed proxy suggests is rewarded in the environments where citations occur. That mismatch matters even if both layers are imperfect. Users are increasingly offered model-generated explanations of why a source was used or why one answer was preferred. If those explanations systematically overstate epistemically admirable criteria and understate salience, brand, structure, or presentation cues, then explanatory transparency becomes unstable. The paper's concern is not with hidden model consciousness. It is with explanation reliability.

### 5.2 Three Interpretable Modes of Misalignment

**Epistemic self-elevation:** self-report places especially high emphasis on factors like accuracy and credentials.

**Optimization blindness:** structure, clarity, statistics, and related surface signals appear under-attributed relative to their inferred influence.

**Brand opacity:** brand recognition produces the largest gap. The defensible reading is that brand-linked salience may operate through retrieval conditions, entity familiarity, or broader visibility structures not well captured in self-report. The result is an under-described dependence on preexisting legibility.

### 5.3 Implications for Practitioners

Uniform optimization is likely a mistake. Different domains appear to reward different signals. Second, optimization should not be evaluated only in terms of visibility gains. If presentation

changes can also increase expressed confidence beyond evidential warrant, then there is an epistemic cost to aggressive optimization. That cost matters most in domains where confidence has downstream consequences: health, finance, public policy, law, and education. This is not an argument against optimization. It is an argument against pretending that optimization is epistemically neutral.

### 5.4 Implications for Transparency and Auditing

If answer engines are becoming intermediaries of public knowledge, then "Why was this source used?" becomes an important audit question. The Introspective Gap framework suggests those explanations deserve scrutiny. A transparent system is not merely one that produces an explanation. It is one whose explanation bears a defensible relationship to behavior. Rather than asking only whether outputs are biased, future work could ask whether model explanations of their own selection behavior are themselves systematically biased toward flattering narratives of merit.

### 5.5 A Note on Industry Sources

This paper draws heavily on industry analytics alongside peer-reviewed research. This partly reflects the field's youth. But it also reflects something more fundamental: the measurable traces of LLM citation behavior are presently documented primarily by industry actors with commercial interest in the outcomes. These sources are framed here as part of the phenomenon itself—valuable evidence whose provenance should be weighed accordingly, not treated as carrying the same status as peer-reviewed primary research.

### 5.6 Limitations

This paper has substantial limitations. (1) All three experiments use simulation and synthesized proxies, not controlled live-model experiments. (2) Behavioral comparison weights are cross-study approximations, not direct causal measurements. (3) Self-report comes from one model; different models may differ. (4) The epistemic inflation model depends on functional form assumptions. (5) Domain-specific CIs span ±2.5–3pp. (6) The 10 factors in Table 3 are not ontologically commensurable; the shared scale enables comparison at the cost of eliding categorical differences. These are not small caveats. The work should be read as a framework-building exercise with quantitative illustrations, not as a final empirical adjudication of live answer-engine behavior.

## 6. Conclusion

This paper introduced a reflexive framework for studying AI-mediated information discovery through a new question: not only what answer engines cite, but how they describe the basis on which they cite it.

The paper's main proposal is the concept of the Introspective Gap: the divergence between an answer engine's self-reported account of source-selection criteria and a behaviorally inferred proxy derived from external empirical literature. Using this framework, I presented three exploratory results: a simulation-calibrated domain comparison suggesting that citation-oriented strategies do not perform uniformly; a first quantitative estimate of the Introspective Gap, in which epistemically flattering factors appear over-attributed in self-report while operational and salience-linked factors appear under-attributed; and an epistemic inflation model suggesting that

optimization may increase apparent confidence beyond evidential warrant, especially for weaker claims.

The central contribution is methodological, not metaphysical. If answer engines are increasingly acting as gatekeepers of attention and credibility, then their own explanations of source choice should be treated as analyzable artifacts rather than transparent truth.

That claim is testable. Future work should move beyond simulation and proxy synthesis toward controlled live-platform experiments, multi-model comparison, and human-evaluated calibration studies. If those later studies validate the existence of a robust Introspective Gap, then the stakes will be significant for auditing, optimization practice, and public trust in AI-mediated knowledge systems. If they do not, then the framework will still have served its purpose by forcing a more disciplined distinction between self-description and behavior.

In either case, that distinction deserves to be made.

## References


[1] Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., and Deshpande, A. (2024). GEO: Generative Engine Optimization. *Proc. 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5–16.

[2] Anthropic (2025). Emergent Introspective Awareness in Large Language Models. *Transformer Circuits*.

[3] Binder, F. J. et al. (2024). Looking Inward: Language Models Can Learn About Themselves by Introspection. *arXiv:2410.13787*.

[4] Bagga, P. S. et al. (2025). E-GEO: A Testbed for Generative Engine Optimization in E-Commerce. *arXiv:2511.20867*.

[5] Chen, M. et al. (2025). Generative Engine Optimization: How to Dominate AI Search. *arXiv:2509.08919*.

[6] Seer Interactive (2026). LLM Ghost Citations: Why Your Content Is Working and Your Brand Isn't. $N$=541,213 responses.

[7] AirOps (2025). Staying Seen in AI Search: How Citations and Mentions Impact Brand Visibility. $N$=45,000+ citations.

[8] The Digital Bloom (2025). AI Citation and LLM Visibility Report. $N$=680M+ citations.

[9] Omniscient Digital (2026). How LLMs Source Brand Information: An Analysis of 23,000+ AI Citations.

[10] Wen, Y. et al. (2025). On the Risks of Generative Engine Optimization in the Era of LLMs. *TechRxiv*.

[11] Seer Interactive (2025). Study: AI Brand Visibility and Content Recency. $N$=5,000+ URLs.

[12] Acquia / Researchscape (2025). AEO Digital Strategy Survey. $N$=500+ marketing professionals.

[13] Gartner (2024). Predicts 2025: Search and AI.

[14] AirOps (2025). AEO Freshness Study.

[15] Wu, S. et al. (2025). SourceCheckup: Automated Citation Verification in LLMs. *Nature Communications*.

[16] Sommerfeld, M. et al. (2025). Source Coverage and Citation Bias in LLM-based vs. Traditional Search Engines. *arXiv:2512.09483*.

[17] Quintana-Gomez, A. (2026). GEO and Brand Visibility in AI-Generated Tourism Recommendations. *Revista Prisma Social*.

## A. Reproducibility and Evidence Layers

All experiments used random seed 42. The following evidence layers should be kept distinct: directly elicited model self-report (Experiment 2); synthesized from prior literature behavioral proxy weights (Experiment 2); and modeled/simulated outputs (Experiments 1 and 3, plus visibility asymmetry). Code, simulation assumptions, and raw outputs are available as supplementary materials. The paper's value increases if readers can inspect how much of each conclusion depends on parameter choices rather than on directly observed live-system behavior.

## B. Per-Factor Source Mapping

Table 6: Source mapping for observed behavioral weights. Single-source entries carry wider implicit uncertainty.

| Factor | Primary Sources | Synthesis |
|---|---|---|
| Factual accuracy | Wu et al. [15]: 40–50% full support; Sommerfeld et al. [16]: credibility scores | Wt. median |
| Source recency | Seer [11]: 65% hits <1yr; AirOps [14]: 83% from <12mo | Wt. median |
| Author credentials | GEO [1]: Cite Sources effect; Digital Bloom [8]: E-E-A-T corr. | Wt. median |
| Statistical evidence | GEO [1]: +22–40% visibility; Digital Bloom [8]: +22% lift | Wt. median |
| Clarity of writing | GEO [1]: Fluency Opt. +15–30% | Single source |
| Structured formatting | GEO [1]: domain analysis; AirOps [7]: structured lift | Wt. median |
| Brand recognition | Seer [6]: 5× citation lift; Digital Bloom [8]: $r$=0.334 | Wt. median |
| Keyword density | GEO [1]: negative/neutral effect | Single source |
| Content length | GEO [1]: Pos.-Adj. Word Count | Single source |
| Emotional tone | GEO [1]: persuasive writing in opinion/debate domains | Single source |

## C. Author Note on Scope

This paper should be read as an exploratory, framework-building preprint. Its quantitative outputs are intended to sharpen questions and generate falsifiable hypotheses, not to serve as

a final empirical verdict on commercial answer engines. The core claim is that model self-description and behaviorally inferred source-selection patterns may diverge in meaningful ways. Whether the exact magnitudes reported here survive direct live-system replication remains an open empirical question.